Detection of Coronary Artery Disease using Logistic Regression

Max Estep

Department of Mathematics and Statistics California State University, Long Beach

I. Introduction

Centers for Disease Control and Prevention (CDC) list Coronary Artery Disease (CAD) as the most common form of heart disease that is responsible for 610,000 deaths in the United States annually[1]. CAD is caused by the buildup of cholesterol deposits in the walls of arteries that supply blood to the heart. Over time, cholesterol deposits, otherwise known as plaque, can cause arteries to partially or completely block blood flow. This process is called atherosclerosis. The restricted blood flow causes discomfort or chest pain called angina. Leaving CAD untreated, causes the heart muscle to weaken over time, which could result in heart failure or an irregular heartbeat, also known as arrhythmia.

Medical professionals measure cholesterol, blood pressure, and sugar levels of the patient before administering medical tests that test for CAD. There are also risk factors that are weighed in, such as being overweight, physical inactivity, unhealthy eating, and smoking tobacco. Aside from common risk factors, medical professionals also suggest a family history of CAD elevates a patient's risk of being diagnosed.

The types of tests used to diagnose symptoms of CAD come in the form of exercise stress tests, electrocardiograms, echocardiograms, cardiac catheterization, and coronary angiograms. The two most common forms of tests are exercise stress tests and electrocardiograms. Exercise stress tests are typically done on a stationary exercise bike or treadmill. The examiner adjusts the training machine with varying difficulty and measures the patient's heart rate as a response to any change. Electrocardiograms measure the electrical rate, activity, and regularity of the patient's heart beat.

Therefore, the purpose of this study will be to employ the results of stress tests to develop candidate logistic regression models that detect CAD in individuals. Factor Analysis of Mixed Data (FAMD) will also be utilized to explore the relationships between predictors; and, to use the generated component scores for further model development. The goal is to search for a model that maintains the sensitivity,specificity, and predictive values needed for diagnosing CAD. The models will be trained using 80% of the data and the remaining 20% will be used for testing. For model validation, the candidate models will be evaluated using Leave-One-Out-Cross-Validation (LOOCV) on the training set and the test set will be used to asses the sensitivity, specificity, and relevant metrics of the fitted models on new data.

II. Data

The data used for this study was collected from the UCI Machine Learning Repository[4] sourced by the University Hospital of Zurich, University Hospital of Base, Hungarian Institute of Cardiology, and V.A. Medical Center of Long Beach. The original data set contained 76 attributes but was reduced to the following 14 attributes.

Name	Description	Type
Age	Age of patient	Cont.
Sex	Sex of patient	Cat.
Cp	Type of chest pain	Cat.
Trestbps	Resting blood pressure	Cont.
Chol	Serum Cholesterol in mg/dl	Cont.
Fbs	Fasting Blood Sugar >120 mg/dl	Cat.
Restecg	Resting Electrocardiographic results	Cat.
Thalach	Maximum heart rate achieved	Cont.
Exang	Exercise induced angina	Cat.
Oldpeak	ST depression induced by exercise rel. to rest	Cont.
Slope	Slope of the peak exercise ST segment	Cat.
Ca	Number of major vessels (0-3) colored by fluoroscopy	Cat.
Thal	Thalassemia level	Cat.
Target	Presence of Heart Disease	Cat.

Table 1: Coronary Artery Disease Variables

The data is based on stress tests performed by 303 individuals recorded at 3 different hospitals. Table 1 lists all of the variables that will be used for model fitting. An in depth data description is included in the attached appendix.

III. Factor Analysis of Mixed Data (FAMD)

Factor Analysis of Mixed Data is a combination of Principal Component Analysis (PCA) and Multiple Correspondence Analysis (MCA) developed by Hill and Smith (1972). There has also been some developments made by Escofier (1979) and Pages (2004) reflected in the R packages *FactoMineR* and *PCAmixdata*. FAMD allows the simultaneous analysis of quantitative and qualitative variables in the same fashion as standard Factor Analysis. There is an in depth analysis of the methodology in Gilbert Saporta's paper *Simultaneous Analysis of Qualitative and Quantitative Data*. The main goal of Factor Analysis is to reduce dimensionality and describe variability among observed correlated variables in terms of latent factors [2].

A. Analyzing the Coronary Artery Disease Data using Factor Analysis of Mixed Data

Just as in standard Factor Analysis, the generated loadings and factor scores can be used for further analysis.

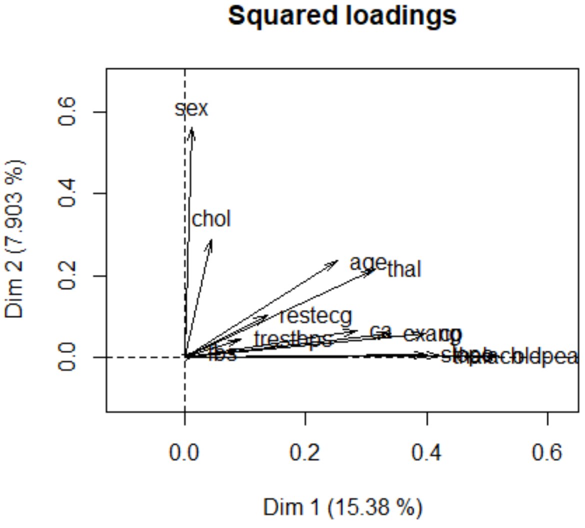

Figure 1: Squared loadings of unrotated components

Figure 1 illustrates the squared loadings of all the predictors and represents their percentage of variation explained in the first two factors. We can also view the squared loadings as a proportion of variation that is derived from the correlation between each factor and variable. With the given representation, interpretation might be troublesome because there is no clear separation between the variables with respect to each dimension. The cumulative amount of variation explained by the first 5 factors only accumulates to 43.02% as shown in Table 2. This amount of variation explained is low, however, the loadings constitute factors that have a relevant interpretation. It appears variables *age, thalach, oldpeak, exang, slope, ca, and thal* correlate most with the first factor. Also, variables *cholesterol* and *sex* are correlated with the second factor. The other components can be discriminated in a similar manner. The first factor could be considered as the stress test response factor because the variables correlated with this component deal exclusively with the patient's physical state after testing. Next, the second factor can be thought of as the *cholesterol and sex* related component. Notice that the first factor involves the variable *age*, this could suggest that the age of the patient may have a noticeable effect on the outcome of the stress test. In other words, it might not be the best practice to have older patients partake in physical related stress tests since other age-related ailments might come into play. If this is the case, testing for symptoms of CAD could be done through the alternative tests discussed earlier if an exercise stress test is unsuitable for the patient [3].

Name	dim 1	dim 2	dim 3	dim 4	dim 5
Age	**0.25**	0.24	0.01	0.02	0.06
Trestbps	0.10	0.06	0.21	0.12	0.01
Chol	0.04	**0.28**	0.02	0.07	0.14
Thalach	**0.42**	0.01	0.08	0.12	0.02
Oldpeak	**0.52**	0.00	0.09	0.03	0.02
Sex	0.01	**0.56**	0.01	0.00	0.01
Cp	**0.40**	0.07	0.29	0.10	0.13
Fbs	0.02	0.02	0.09	0.10	**0.30**
Restecg	0.13	0.11	0.06	**0.36**	0.02
Exang	**0.34**	0.06	0.07	0.00	0.01
Slope	**0.40**	0.01	0.39	0.10	0.01
Ca	**0.26**	0.02	0.19	0.24	0.22
Thal	**0.32**	0.21	0.08	0.09	0.27
Eigenvalue	3.23	1.66	1.58	1.34	1.22
Proportion	15.38%	7.9%	7.5%	6.39%	5.83%
Cumulative	15.38%	23.28%	30.79%	37.19%	43.02%

Table 2: Squared loadings of unrotated components

Name	dim 1	dim 2	dim 3	dim 4	dim 5
Age	0.15	**0.34**	-	-	-
Trestbps	0.06	0.09	-	-	-
Chol	0.01	**0.32**	-	-	-
Thalach	**0.38**	0.05	-	-	-
Oldpeak	**0.51**	0.01	-	-	-
Sex	0.07	**0.49**	-	-	-
Cp	**0.43**	0.05	-	-	-
Fbs	0.01	0.02	-	-	-
Restecg	0.09	0.15	-	-	-
Exang	**0.39**	0.02	-	-	-
Slope	**0.39**	0.02	-	-	-
Ca	**0.24**	0.04	-	-	-
Thal	**0.42**	0.11	-	-	-
Eigenvalue	3.16	1.73	-	-	-
Proportion	15.7%	8.52%	-	-	-
Cumulative	15.7%	24.22%	-	-	-

Table 3: Squared loadings of first two rotated components

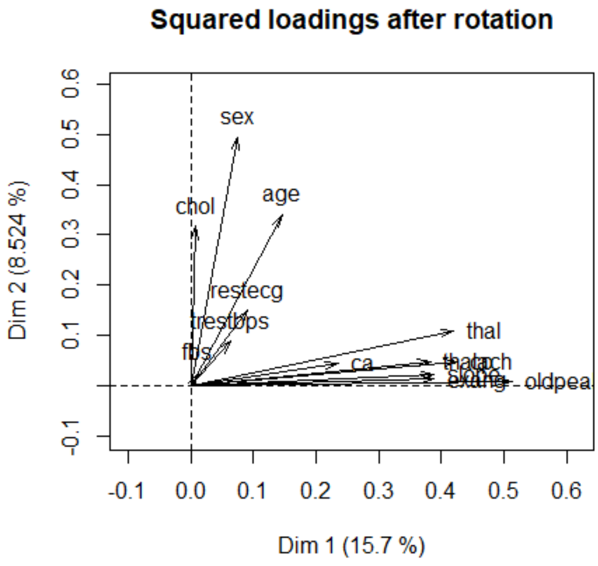

Figure 2: Squared loadings after rotation.

We do have a reasonable way of interpreting the squared loadings, however, we could possibly improve these interpretations by rotating the components. Table 3 contains the squared loadings for the first two rotated components and Figure 2 displays the clear separation we sought after between components. After rotation, the interpretation of the squared loadings become more sensible. The first factor's interpretation remains the same after rotation. However, the second factor can now be thought of as the factor corresponding to the *sex, age, and cholesterol* of the patient.This is quite an interesting factor because according to the CDC men are more likely to be diagnosed with CAD at a younger age than women; contrarily, around the same number of men and women die each year to CAD[1]. Nonetheless, even though the rotated components result in a clearer factor interpretation, we are still going fit logistic regression models using the rotated and unrotated factor components for general comparisons.

IV. Model Development

The first candidate model we are going to consider for the analysis of the CAD data is a logistic regression model strictly using the original data. For logistic regression, the response variable can take on a dichotomous role, taking on only two levels, or more than two levels which extends beyond dichotomy. In the case of the CAD data, the response variable target is dichotomous, where the label 0 indicates the absence of coronary artery disease and 1 represents the presence of coronary artery disease. The logistic regression model with k predictors has the form:

$$logit(p) = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

When a categorical predictor with n factor levels is present, n-1 dummy variables will be utilized for n-1 factor levels with the remaining level being left as the reference variable. For example, if variable X_k has 3 levels then there would be coefficients for 2 of the factor levels included in the model:

$$\beta_{1k} X_{1k} + \beta_{2k} X_{2k}$$

These coefficients would describe (all other things held constant) how the log odds of the desired event occurring changes with respect to the corresponding factor level being present.

For each of the final models given in this section, we will check model assumptions and adequacy based on the following:

1. Linearity: Predictors are assumed to be linearly related to the log-odds.

2. Collinearity: Predictors are assumed to not be linearly dependent.

3. Residuals: The binned residuals are approximately normally distributed around 0 with constant variance.

A. Original Data Logistic Regression Model (ODLRM)

Stepwise regression, backward elimination, and the forward selection methods were applied to a fully fitted logistic regression model with all the training data for variable screening. Of these three methods, stepwise regression and backward elimination produced the same model with an AIC of 163.5 which is lower than that of the forward selection method (171.43). This model summary is attached in the appendix with the label *1st Reduced ODLRM*.

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	0.426131	4.366156	0.098	0.922251
Age	0.018790	0.031132	0.604	0.546135
Sex [1]	-1.981701	0.705544	-2.809	**0.004973**
Cp [1]	1.21916	0.755597	1.614	0.106618
Cp [2]	2.097572	0.617911	3.394	**0.000688**
Cp [3]	3.550737	0.955114	3.718	**0.000201**
Trestbps	-0.028436	0.015307	-1.858	**0.063206**
Chol	0.001853	0.005767	0.321	0.747988
Fbs [1]	0.161477	0.693562	0.233	0.815899
Restecg [1]	0.708374	0.512277	1.383	0.166728
Restecg [2]	-0.351662	2.661695	-0.132	0.894890
Thalach	0.02393	0.014757	1.763	**0.077840**
Exang [1]	-1.379884	0.554346	-2.489	**0.012803**
Oldpeak	-0.586872	0.317688	-1.847	**0.064701**
Slope [1]	-1.775038	1.168946	-1.518	0.128890
Slope [2]	-0.619559	1.301506	-1.518	0.634051
Ca [1]	-2.500550	0.629655	-3.971	**7.15e-05**
Ca [2]	-4.093382	1.023010	-4.001	**6.30e-05**
Ca [3]	-1.979068	1.023010	-2.065	**0.038964**
Ca [4]	-0.090318	3.242575	-0.028	0.977779
Thal [1]	1.499360	3.088628	0.485	0.627361
Thal [2]	1.815794	2.920972	0.622	0.534178
Thal [3]	0.550662	2.933810	0.188	0.851116

Table 4: Full Original Data Logistic Regression Model

The output in Table 4 shows the variables *age, chol, fbs, restecg, slope, and thal* as being insignificant. We also see in the output that one level of the *Ca* (Ca [4]) and *Cp* (Cp [1]) variables are not significant. Even though they are not significant we are not going to remove these factor levels because they would change the entire model. For categorical predictors, either the whole variable is dropped from the model or left in.

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	1.53834	2.41465	0.637	0.524068
Sex [1]	-2.13842	0.56640	-3.775	**0.000160**
Cp [1]	1.54779	0.718522	2.155	**0.031159**
Cp [2]	1.97481	0.54524	3.622	**0.000292**
Cp [3]	3.46963	0.90072	3.852	**0.000117**
Trestbps	-0.03012	0.01368	-2.203	**0.027618**
Thalach	0.03399	0.01254	2.713	**0.006673**
Exang [1]	-1.52686	0.52118	-2.930	**0.003394**
Oldpeak	-0.60696	0.24835	-2.444	**0.014526**
Ca [1]	-2.20963	0.55144	-4.007	**6.15e-05**
Ca [2]	-3.39810	0.81422	-4.173	**3.00e-05**
Ca [3]	-1.88550	0.86617	-2.177	**0.029494**
Ca [4]	-0.25966	1.85987	-0.140	0.888968

Table 5: 2nd Reduced Original Data Logistic Regression Model

Removing the *slope* variable from the model was considered because neither of the two levels that were used as dummy variables were found to be significant. Table 5 summarizes the changes to the model after removing insignificant variables. The removal of the *slope* variable increased the AIC to 166.43.

1. Linearity of Independent Variables and Log-Odds

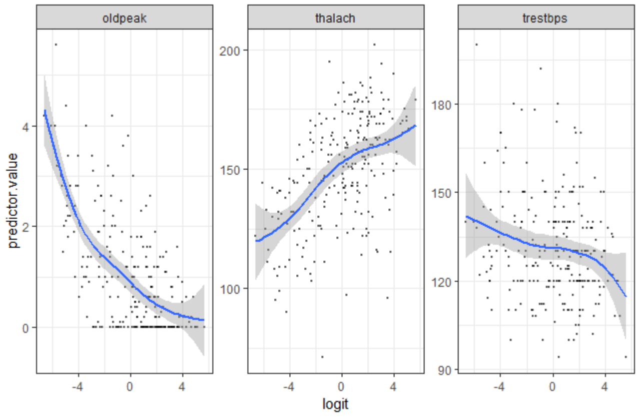

Figure 3: Linearity Check for *2nd reduced ODLRM*

Figure 3 illustrates an approximate linear relationship between continuous predictors and log-odds.

2. Multicollinearity

Variable	VIF	Variable	VIF	Variable	VIF
sex	1.199452	cp	1.683560	trestbps	1.086868
thalach	1.243546	exang	1.161053	oldpeak	1.386450
ca	1.374835				

Table 6: *2nd Reduced ODLRM* VIF

Since all of the variance inflation factors (VIF) are less than 3, it is safe to assume multicollinearity is absent from the model.

3. Residuals

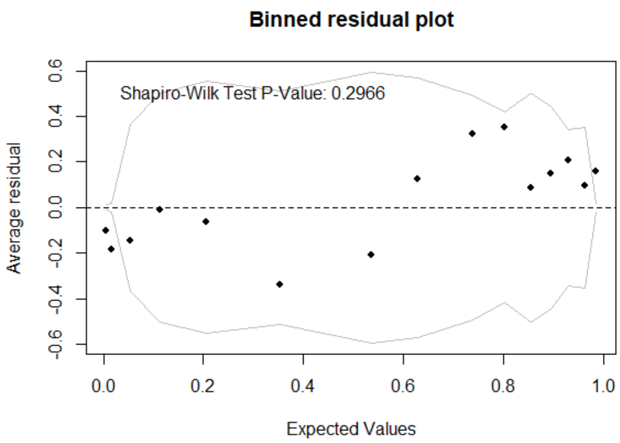

Figure 4: Binned Residuals for *2nd Reduced ODLRM*

Based on the Shapiro-Wilk Test, we can assume the binned residuals are normally distributed for the *2nd Reduced ODLRM*.

B. Unrotated Factor Score Logistic Regression Model (UFSLRM)

Since the factor score models deal with a smaller amount of predictors, we can easily screen variables without the need of using screening methods as we had done in the *ODLRM* section.

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	0.1945	0.1875	1.037	0.29980
F [1]	-1.2392	0.1586	-7.813	**5.57e-15**
F [2]	0.5806	0.1463	3.968	**7.25e-05**
F [3]	0.4826	0.1432	3.371	**0.00075**
F [4]	-0.3264	0.1540	-2.119	**0.03409**
F [5]	0.0356	0.1664	0.214	0.83058

Table 7: Full Unrotated Factor Score Logistic Regression Model

Based on Table 7, we can see the only insignificant factor score predictor is *F[5]*. This predictor represents the dichotomous variable *Fbs* that defines whether a patient's fasting blood sugar is below or above 120 mg/dl.

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	0.1936	0.1875	1.033	0.301736
F [1]	-1.2389	0.1586	-7.811	**5.69e-15**
F [2]	0.5794	0.1459	3.970	**7.18e-05**
F [3]	0.4808	0.1426	3.372	**0.000746**
F [4]	-0.3269	0.1540	-2.122	**0.033798**

Table 8: Reduced Unrotated Factor Score Logistic Regression Model

After reducing the model to only the significant predictors, as seen in Table 8, the model's AIC decreases from 193.39 to 191.44.

1. Linearity of Independent Variables and Log-Odds

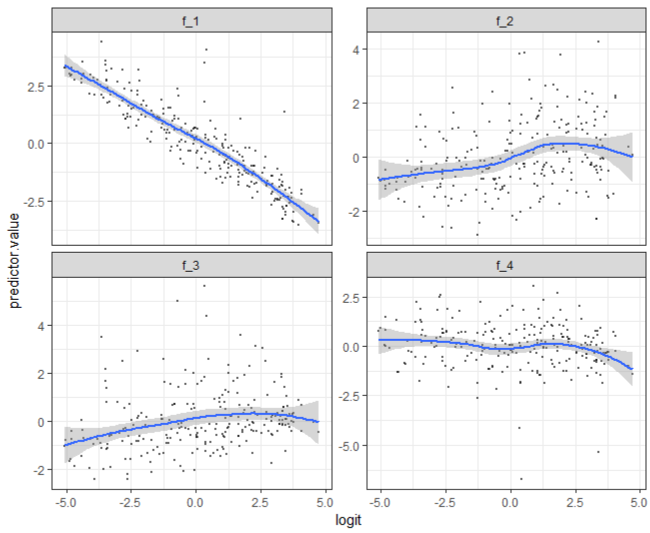

Figure 5: Linearity Check for *Reduced UFSLRM*

Figure 5 illustrates an approximate linear relationship between continuous predictors and log-odds.

2. Multicollinearity

Variable	VIF	Variable	VIF	Variable	VIF
F[1]	1.203903	F[2]	1.068819	F[3]	1.082021
F[4]	1.068502				

Table 9: *Reduced UFSLRM* VIF

We can assume the factor score predictors are not collinear since their variation inflation factors (VIF) are below 3.

3. Residuals

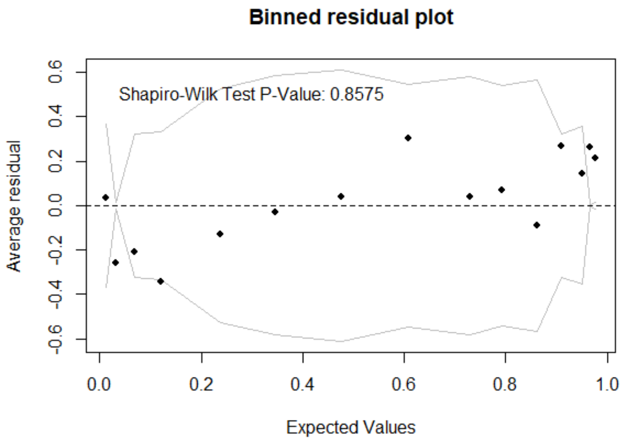

Figure 6: Binned Residuals for *Reduced UFSLRM*

Based on the Shapiro-Wilk Test, we can assume the binned residuals are normally distributed for the *Reduced UFSLRM*.

C. Rotated Factor Score Logistic Regression Model (RFSLRM)

We base the rotated factor score models on the rotation of the first two factors derived from the factor analysis in the *FAMD* section.

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	0.2158	0.1786	1.208	0.227
F [1]	-1.2062	0.1468	-8.216	**2e-16**
F [2]	0.1949	0.1351	1.443	0.149

Table 10: Full Rotated Factor Score Logistic Regression Model

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	0.2123	0.1772	1.198	0.231
F [1]	-1.2057	0.1471	-8.197	**2.47e-16**

Table 11: Reduced Rotated Factor Score Logistic Regression Model

The only factor score predictor that remains significant after reducing the model is *F[1]*. Despite the model reduction, the AIC has marginally increased from 203.9 to 204.03.

1. Linearity of Independent Variables and Log-Odds

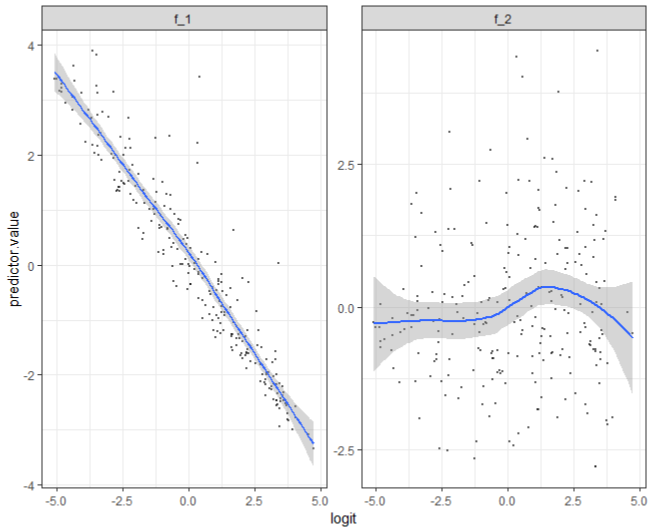

Figure 7: Linearity Check for *Full RFSLRM*.

There is clearly an approximate linear relationship between the predictors and log-odds.

2. Multicollinearity

Variable	VIF	Variable	VIF
F[1]	1.004713	F[2]	1.004713

Table 12: 2nd Reduced Original Data Logistic Regression Model

We can assume the factor score predictors are not collinear since their variation inflation factors (VIF) are below 3.

3. Residuals

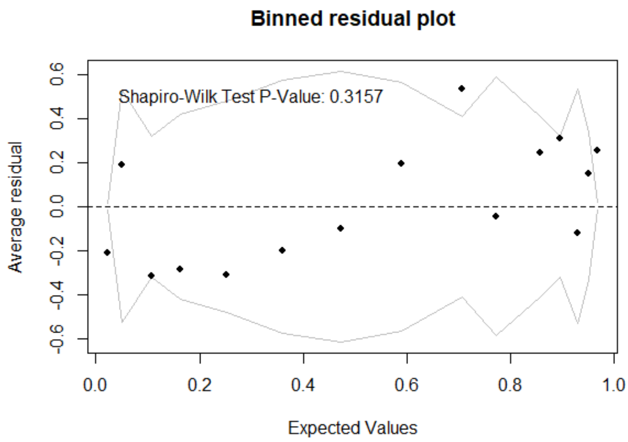

Figure 8: Binned Residuals for *Full RFSLRM*

Based on the Shapiro-Wilk Test, we can assume the binned residuals are normally distributed for the *Full RFSLRM*.

V. Model Evaluation

We are going to focus on a couple key metrics in this section to evaluate the performance of each model defined in the *Model Development* section. Leave-one-out-cross-validation (LOOCV) was used to check if the models defined previously are consistent with the proposed reduced model.

Model	CV Err.	AIC	Spec.	Sens.	NPV	PPV	ACC
ODLRM							
Reduced [1]	0.103	163.53	0.67	0.88	0.82	0.76	0.78
Reduced [2]	0.109	166.43	0.67	0.79	0.72	0.74	0.73
UFSLRM							
Full	0.247	193.39	0.85	0.91	0.88	0.89	0.89
Reduced	0.244	191.44	0.85	0.91	0.88	0.89	0.89
RFSLRM							
Full	0.157	203.90	0.93	0.88	0.86	0.94	0.90
Reduced	0.157	204.03	0.93	0.88	0.86	0.94	0.90

Table 13: Model Evaluation Metrics

Next, using the test data, the given models were used to calculate the specificity, sensitivity, and predictive values. Since medical studies are mainly concerned with Type II error, or the false negative rate, we want to select a model that minimizes this metric. Based on Table 10, the *Reduced UFSLRM* has a false negative rate (FNR) of $\sim 9\%$, *Full RFSLRM* has a $\sim 12\%$ FNR, and *Reduced [1] ODLRM* has the same FNR as the *Reduced UFSLRM*. Overall the *Full RFSLRM* has the highest accuracy of $\sim 90.2\%$ using the test data.

A. ODLRM Interpretation

As an example, we will use the 2nd reduced model to interpret a selection of significant coefficients to inspect how they influence the odds of being diagnosed with CAD. Referencing Table 5 on page [3], we can see the variables *sex, cp, thalach, exang, and ca* are a collection of significant variables. Out of these 5 variables we will interpret *sex and trestbps* as examples to show how to interpret a qualitative and quantitative predictor.

1) Interpretation of Sex[1] Coefficient: For a one-unit increase in *Sex[1]*, from female to male, we expect a 2.13842 decrease in the log-odds of detecting CAD.

2) Interpretation of Thalach Coefficient: For a one-unit increase in *Thalach*, we expect a 0.03399 increase in the log-odds of detecting CAD.

B. UFSLRM and RFSLRM Interpretation

We will use the Full *RFSLRM* as an example to explain how to interpret coefficients corresponding with factor scores.

1) Interpretation of F[1] Coefficient: For a one-unit increase in *F[1]*, the factor score predictor representing the stress test response, we expect a 1.2057 decrease in the log-odds of detecting CAD. If a patient has a low stress test response score, the log-odds of them being diagnosed increases and a high stress test response score results in a lower log-odds of being diagnosed.

2) Interpretation of F[2] Coefficient: For a one-unit increase in *F[2]*, the factor score predictor representing *age, sex and cholesterol (ASC)*, we expect a 0.1949 increase in the log-odds of detecting CAD. If a patient has a low ASC score, the log-odds of them being diagnosed decrease and a high ASC score corresponds to a higher log-odds of being diagnosed.

VI. Conclusion

After going through the process of conducting a factor analysis of mixed data, model development and evaluation, we produced a selection of models that have modest sensitivity, specificity, and predictive values. Out of all the models evaluated, the unrotated (*UFSLRM*) and rotated (*RFLSRM*) factor score models yielded the highest accuracy (90.2%) and related metrics when compared against the original (*ODLRM*) data models. Of the models evaluated, the reduced *UFSLRM* has a $\sim 9\%$ FNR followed the full *RFSLRM* with a $\sim 12\%$ FNR. One of the more significant benefits of using the factor score models is their predictability and ease of interpretation. The factor score models also have meaningful factors that coincide with the discussions in the *Introduction* section. Likewise, *ODLRM's* have ease of interpretation with the added flexibility of being able to conduct tests on individual coefficients for specific levels.

References

[1] Centers for Disease Control and Prevention. 2017. Heart Disease. Retrieved from http://www.cdc.gov/heartdisease/

[2] Marie Chavent,Vanessa Kuentz-Simonet,Amaury Labenne ,Jerome Saracco. Multivariate Analysis of Mixed Data: The R Package PCAmixdata., 2014,Retrieved April 3, 2019, from the arXiv database.

[3] Deanna L. Huggett, Denise M. Connelly, Tom J. Overend, Maximal Aerobic Capacity Testing of Older Adults: A Critical Review, The Journals of Gerontology: Series A, Volume 60, Issue 1, January 2005, Pages 5766, https://doi.org/10.1093/gerona/60.1.57

[4] Dua, D. and Graff, C. 2019. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

[5] Johnson, Richard A, and Dean W. Wichern. Applied Multivariate Statistical Analysis. Englewood Cliffs, N.J: Prentice Hall, 1992. Print

Appendix

A. 1st Reduced ODLRM

Coefficient	Estimate	Std. Err.	Z Value	Pr(> |Z|)
Intercept	3.82553	2.77595	1.378	0.168173
Sex [1]	-2.34996	0.60122	-3.908	**9.20e-05**
Cp [1]	1.35271	0.73051	1.852	**0.064065**
Cp [2]	2.08275	0.57566	3.618	**0.000297**
Cp [3]	3.55633	0.89270	3.984	**6.78e-05**
Trestbps	-0.02755	0.01403	-1.964	**0.049530**
Thalach	0.02393	0.01334	1.793	**0.072981**
Exang [1]	-1.52915	0.53674	-2.849	**0.004386**
Oldpeak	-0.58126	0.28059	-2.072	**0.038304**
Slope [1]	-1.57407	1.04488	-1.506	0.131947
Slope [2]	-0.31603	1.13185	-0.279	0.780082
Ca [1]	-2.44450	0.58376	-4.188	**2.82e-05**
Ca [2]	-3.76703	0.90308	-4.171	**3.03e-05**
Ca [3]	-2.03659	0.91427	-2.228	**0.025910**
Ca [4]	-0.23614	2.30402	-0.102	0.918366

B. Data Description

Name	Description	Type
Age	Age of patient	Cont.
Sex	Sex of patient	Cat.
	Female = 0	
	Male = 1	
Cp	Type of chest pain	Cat.
	Typical Angina = 0	
	Atypical Angina = 1	
	Non-Anginal Pain = 2	
	Asymptomatic = 3	
Trestbps	Resting blood pressure	Cont.
Chol	Serum Cholesterol in mg/dl	Cont.
Fbs	Fasting Blood Sugar $>$120 mg/dl	Cat.
	FBS $<$ 120 mg/dl = 0	
	FBS $>$ 120 mg/dl = 1	
Restecg	Resting Electrocardiographic results	Cat.
	Normal = 0	
	ST-T Wave Abnormality = 1	
	Pos. Left Vent. Hyper. = 2	
Thalach	Maximum heart rate achieved	Cont.
Exang	Exercise induced angina	Cat.
	False = 0	
	True = 1	
Oldpeak	ST depression induced by exercise rel. to rest	Cont.
Slope	Slope of the peak exercise ST segment	Cat.
	Upsloping = 0	
	Flat = 1	
	Downsloping = 2	
Ca	Number of major vessels (0-3) colored by fluoroscopy	Cat.
Thal	Thalassemia level	Cat.
	Normal = 1	
	Fixed Defect = 2	
	Reversable Defect = 3	
Target	Presence of Heart Disease	Cat.
	No presence of heart disease = 0	
	Presence of heart disease = 1	